A Multinomial Model for the Quality Control of Colony Counting Procedures

by B. Voss*, S. Dahms•, J. Kunert* and H. Weiss•

* Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

• Institute of Biometry and Data Processing, Free University of Berlin, D- 14195 Berlin, Germany

Summary

The so-called good-laboratory-practice (GLP) test provides an experimental design and appropriate statistical analysis for the problem of analyst performance assessment in microbiological laboratories. For a given sample material multiple dilution series are generated yielding colony counts from several dilution levels. Statistical evaluation is based on the assumption of Poisson-distributed colony forming units. In this paper a new model based on conditional binomial and multinomial distributions is presented and it is shown how it is related to the standard model which assumes Poisson-distributed colony counts. The effects of common working errors on the statistical evaluation of the GLP-test are investigated.

Keywords: standardized analytical methods, Good-laboratory-practice-test, Poisson model, multinomial model, quality control

1 Introduction

Standardized colony counting procedures are essential for the microbiological quality control of milk and milk products. The standardisation of an analytical technique requires the estimation of precision values like repeatability within laboratories and its reproducibility between different laboratories according to (DIN ISO 5725, 1988). These values are estimated from collaborative studies with so-called routine laboratories. However, prerequisite for participation in a collaborative study of this kind is a thorough demonstration that the laboratory staff involved performs the analytical technique in an acceptable working standard. As a proficiency testing scheme for the pour-plate technique Weiss, Nimeläe and Arndt (1991) have introduced the so-called good-laboratory-practice (GLP) test which includes an experimental design and appropriate statistical analysis. It provides both, a method for estimating germ concentration and a quality assurance system for checking whether the measurement technique in a laboratory was applied according to an acceptable standard (IDF, 1994; BGA, 1991). By observing multiple dilution series and inoculating several plates from the same dilution level within laboratory errors can be detected and their sources identified.

The main assumption underlying the statistical analysis of the GLP-test is that faultless performance of the analytical technique procedures without any working errors results in Poisson-distributed colony counts. Dahms (1996) has extensively studied the evaluation strategy for the GLP-test. Using simulation models reflecting the different steps of the experiment, she investigated effects of ideal work and several combinations of working errors on the distribution of test statistics used in the GLP decision scheme. She reported that even in the case of faultless work variability of counts is higher than expected in the Poisson model. Therefore, Dahms criticised the homogeneity test as not adequate for assessing analyst performance and questioned the appropriateness of this test in the context of a GLP evaluation.

In this article a new model for describing the distribution of colony counts in the GLP-test is proposed which is based on the multinomial distribution. It is shown how this model is connected to the Poisson model of Weiss et al.. The effects of the working errors explored in the simulation studies of Dahms are investigated within the context of the multinomial model and a

new assessment for the performance of the GLP-test is derived.
After this introduction, Section 2 reviews the experimental design and the statistical analysis of the GLP-test together with the Poisson model. Section 3 introduces the multinomial model, while simulated working qualities and results of the performed simulations are described in Section 4.

2 Experimental design and statistical analysis of the GLP-test

The design and the analysis of the GLP-test is described extensively in the DIN ISO/CD 14461 (1996). For convenience, we give a short review here.
First the sample material is prepared, homogenised and diluted to a suitable working density. The experimental design consists of four sequential dilution series. In each series the sample material is diluted 12 times each time by factor two. On each level in each series three plates are inoculated with the sample material. This is performed sequentially, until the twelfth level is reached. Then the next series is started. The number of colonies grown at the plates after some time of incubation are counted and analysed then. See Figure 1 for a graphical presentation of the experimental design.
If the experiment is done accurately, one would expect the average number of colonies per millilitre to decrease by its half with every dilution step. Before counting the grown colonies the plates should be coded and randomised to prevent an assimilation of counting results to each other by the analyst. Adequacy of collected data should be checked before statistical evaluation.
Following recommendations of Baumgart (1986) only those dilution steps are suitable for evaluation for which the arithmetic mean of counts on the three parallel plates is less than 300 for all four dilution series. Additionally, for

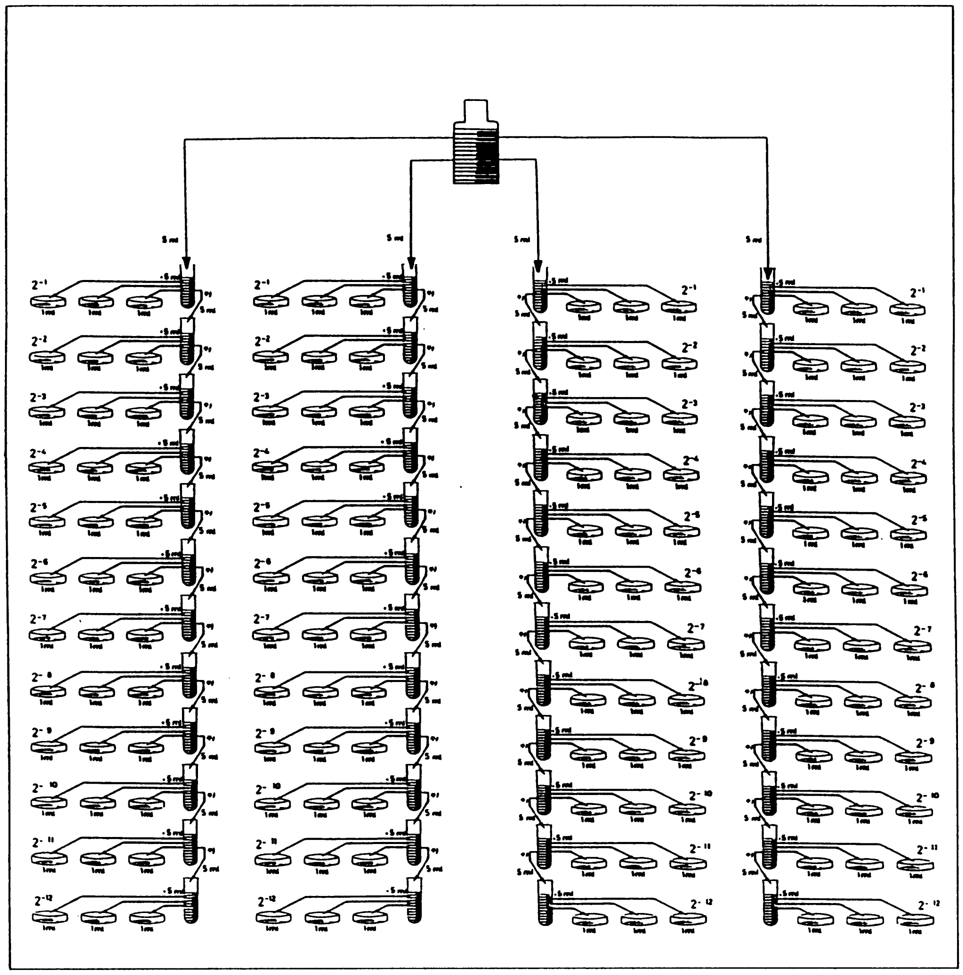

Figure 1: Experimental design of the GLP-test.

an evaluable dilution step the expected average count over its twelve plates must be at least five. The expectation can be estimated by calculating the Farmiloe-estimator (Farmiloe et al., 1954), described later in this text, over all twelve dilution steps. Usually, six evaluable dilution steps covering 72 plates can be expected to be suitable for analysis. If less than five dilution levels are left for evaluation, the experiment should be repeated.

The analysis of the collected data is based on the assumption that the colony numbers on the plates follow a Poisson distribution with a parameter λ describing the average number of colony forming units in one millilitre prediluted sample material. In the experimental design the concentration is reduced by its half with every dilution step. If the techniques of diluting, inoculating the plates and counting are performed without any errors and the dilution error is negligible, the average number of colony forming units in one millilitre of sample material is reduced by its half with each dilution step also. Therefore, the statistical model describes the counted number of colonies on a single plate by a Poisson distribution with parameter $\frac{1}{2^i}\lambda$ depending on the dilution level:

$$X_{ijk} \sim P_{\frac{1}{2^i}\lambda}.$$

Here, X_{ijk} denotes the counted number of colonies on the ith evaluable dilution step in the jth series of the kth parallel plate.
Statistical analysis of the observed colony counts is done in two steps. The first one is based on two likelihood-ratio-tests performing a qualitative evaluation of the noise in the data. First the hypothesis of independence of the parallel plates is tested against the alternative that variability of counts between parallel plates is smaller than expected in the Poisson model ("underdispersion"). If this hypothesis is rejected, the experiment has to be redone. Otherwise the test of overall homogeneity of counts is carried out. If this hypothesis cannot be rejected, the noise in the data can be described by the Poisson model. This indicates that the laboratory counts independently. If it is rejected, the noise is bigger than one would expect for a random error of a Poisson- distributed variable. A large variability indicates problems with performing the standardised method in at least one step. Therefore, as a second step Weiss et al. suggest an analysis of variance with random effects to get more information about the additional variation of the counts. This second part finally determines whether the analyst is able to use perform the microbiological method in a satisfactory way.

As described above, the first test, the test for internal homogeneity checks if the colony counts of the parallel plates show less than the expected variation:

$\mathcal{H}_0$: In every series and on every dilution level the variability of colony counts between parallel plates can be described with the variance of a Poisson-distribution,

$\mathcal{H}_A$: variation of counts is too small.

The likelihood-ratio statistic for this hypothesis

$$G_S^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk} ln \frac{X_{ijk}}{\bar{X}_{ij.}}$$

follows approximately a chi-square distribution with $IJ(K-1)$ degrees of freedom. Here, I denotes the number of evaluable dilution steps and $\bar{X}_{ij.}$ indicates averaging over the $K = 3$ plates in the ith dilution step in the jth series with $J = 4$. The hypothesis $\mathcal{H}_0$ is rejected if

$$G_S^2 < \chi^2_{IJ(K-1),0.005}.$$

If the nullhypothesis is rejected, the variation of the colony counts of the three parallel plates is too small for Poisson-distributed random variables. This may be caused by insufficient randomisation of the plates and assimilation of counts from parallel plates to each other while counting. So independence of the colony counts can not be assumed and the experiment has to be redone. If the nullhypothesis is not rejected, overall homogeneity is tested by comparing the colony counts with the values expected under the assumptions of the Poisson model. The expected counts of colonies $\bar{X}_F$ can be estimated by the Farmiloe-estimator (Farmiloe et al., 1954):

$$\bar{X}_F = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk}}{\sum_{i=1}^{I} JK2^{1-i}}.$$

This formula relates the colony counts to the volumina of the sample material, that plates have been inoculated with and estimates the number of colony forming units (c.f.u.) in a volume equal to that volume of sample material used on the first evaluable dilution step. The expected counts for the ith dilution level can be estimated by

$$\bar{X}_{F_i} = 2^{1-i}\bar{X}_F.$$

The test of overall homogeneity is used to decide whether the colony counts can be described by the Poisson model:

$\mathcal{H}_0$: The overall variation of the colony counts can be described by the Poisson model ,

$\mathcal{H}_A$: additional variation is present in the data.

The likelihood-ratio statistic for this hypothesis

$$G_A^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} X_{ijk} ln \frac{X_{ijk}}{\bar{X}_{F_i}}$$

follows approximately a chi-square distribution with $IJK - 1$ degrees of freedom and $\mathcal{H}_0$ is rejected if

$$G_A^2 > \chi^2_{IJK-1,0.99}.$$

If the hypothesis of overall homogeneity is not rejected, it can be assumed that no additional variation caused by an incorrect application of the laboratory technique is present in the data. Therefore, the analyst performance is acceptable and the laboratory works in a satisfactory way. The statistical analysis can be concluded here.

If the hypothesis of overall homogeneity is rejected, statistical analysis proceeds with an analysis of variance which tries to split up the observed variability into several components to identify possible sources of variation. The observed data are transformed to stabilise the variance and eliminate the effects of the dilution levels:

$$Y_{ijk} = \sqrt{X_{ijk}} - \sqrt{\bar{X}_{F_i}}.$$

The analysis is based on a hierarchical model with random effects for the dilution series R_j and dilution level $S_{i(j)}$ within the series. The variance of the plates $P_{k(i(j))}$ comprises the random error of the data and also variation induced by problems with counting colonies:

$$Y_{ijk} = \mu + R_j + S_{i(j)} + P_{k(i(j))}.$$

If the data follow the Poisson model and there is no additional variance due to dilution series, dilution steps or counting of colonies, an application of the delta method (Serfling, 1980, p. 180) shows that the variance of the Y_{ijk} is approximately 0.25. It is assumed that variation caused by dilution levels within the sequential series can be explained by methodical errors when diluting suspensions. Effects for the series may be due to incomplete homogenisation of the material and a larger variance for the parallel plates indicates mainly problems with counting the colonies. The analyst performance is qualified as unacceptable when the estimated total variance of the transformed data is greater than 1.

3 New Statistical Model

The Poisson model is based on the assumption that samples are taken out parallel and that the numbers of colony forming units (c.f.u.) in the different samples are independent. This includes that an infinite number of c.f.u. is assumed from which any given number can be taken out. In laboratory practice the number of c.f.u. in the initial sample is determined by preparing the material for the experiment. Therefore, the number of c.f.u. in each sample depends on the number of c.f.u. in the initial sample and also on the samples taken out before. Thus, the assumption of independent samples seems to be critical.

The experimental design prescribes that for each dilution series an initial sample is obtained from the prediluted material. Therefore, the colony counts in the different dilution steps of a particular series are dependent on each other because they were all derived from the same initial sample. On the other hand, colony counts on the plates of the same dilution step within a series are negatively correlated. The higher the number of c.f.u. on one of the three plates is, the fewer are left for the other two.

For the new model a fixed number n of c.f.u. in the prediluted material is assumed, which may be considered as a realisation of a Poisson-distributed variable Z with $Z \sim P_{\lambda V}$. Each germ is sampled with the same probability which is the ratio of sample volume to the volume of the material where it is obtained from. So the probability for taking out a specific number of c.f.u. with one of the initial samples can be described by a binomial distribution.

In the GLP-experiment the four initial samples for starting the dilution series are obtained one after the other. Employing the assumption of a fixed number of c.f.u. in the prepared material the distribution of c.f.u. in the four initial samples and the number of c.f.u. remaining in the sample material can be described by conditional binomial distributions. Let $N_j^{(1)}, j = 1, 2, 3, 4$ denote the number of c.f.u. in the jth initial sample and $n_j^{(1)}$ their realisation, then we assume:

$$\begin{aligned}
N_1^{(1)} &\sim Bin(n, p_{N_1^{(1)}}),\\
N_2^{(1)}\,|\,N^{(1)} = n_1^{(1)} &\sim Bin(n - n_1^{(1)}, p_{N_2^{(1)}}),\\
N_3^{(1)} \mid N_1^{(1)} = n_1^{(1)}, N_2^{(1)} = n_2^{(1)} &\sim Bin(n - n_1^{(1)} - n_2^{(1)}, p_{N_3^{(1)}}),\\
N_4^{(1)}\,|\,N_1^{(1)} = n_1^{(1)}, N_2^{(1)} = n_2^{(1)}, N_3^{(1)} = n_3^{(1)} &\\
&\sim Bin(n - n_1^{(1)} - n_2^{(1)} - n_3^{(1)}, p_{N_4^{(1)}}).
\end{aligned}$$

The probabilities $p_{N_j^{(1)}}$ depend on the volumes of the samples v_N (5 ml for GLP-test) and of the remaining prediluted material where the sample is obtained from:

$$p_{N_j^{(1)}} = \frac{v_N}{V - (j-1)v_N}.$$

It follows that the joint distribution of these conditional binomial distributions is a multinomial distribution (Schlittgen, 1996; Kotz and Johnson, 1981):

$$(N_1^{(1)}, N_2^{(1)}, N_3^{(1)}, N_4^{(1)}, N_{Rest}^{(0)}) \sim Mult(n, p_R, p_R, p_R, p_R, 1 - 4p_R)$$

with probabilities $p_R = \frac{v_N}{V}$.

From each of the four initial samples three samples are taken to inoculate the three parallel plates and one sample is obtained to provide material for the second dilution step. If $N_j^{(2)}$ denotes the number of c.f.u. in this sample for the jth series and $N_{jRest}^{(1)}$ indicates the number of c.f.u. remaining in the jth initial sample after the four samples to be taken from it have been obtained, then we assume that this can be described by conditional binomial distributions also:

$$\begin{aligned}
X_{1j1} &\sim Bin(n_j^{(1)}, p_{x_1}), \\
X_{1j2} \mid X_{1j1} = x_{1j1} &\sim Bin(n_j^{(1)} - x_{1j1}, p_{x_2}), \\
X_{1j3} \mid X_{1j1} = x_{1j1}, X_{1j2} = x_{1j2} &\sim Bin(n_j^{(1)} - x_{1j1} - x_{1j2}, p_{x_3}), \\
N_j^{(2)} \mid X_{1j1} = x_{1j1}, X_{1j2} = x_{1j2}, X_{1j3} = x_{1j3} & \\
&\sim Bin(n_j^{(1)} - x_{1j1} - x_{1j2} - x_{1j3}, p_N),
\end{aligned}$$

with probabilities depending on the volume v_x of the samples used for inoculating the plates (1 ml for GLP) and the volume of the diluted material on each step V_{Step} (10 ml):

$$p_{x_k} = \frac{v_x}{V_{Step} - (k-1)v_x} \quad \text{and} \quad p_N = \frac{v_N}{V_{Step} - 3v_x}.$$

The joint distribution of these counts then again is a multinomial distribution:

$$(X_{1j1}, X_{1j2}, X_{1j3}, N_j^{(2)}, N_{j_{Rest}}^{(1)}) \sim Mult(n_j^{(1)}, p_x, p_x, p_x, p_S, 1 - 3p_x - p_S).$$

with

$$p_x = \frac{v_x}{V_{Step}} \quad \text{and} \quad p_S = \frac{v_N}{V_{Step}}.$$

The numbers of c.f.u. in the samples of the later dilution steps can be described in a similar way.

Combining the distributions introduced above to an overall joint distribution, the numbers of c.f.u. on the different dilution steps and plates can be described with a multinomial distribution:

$$(X_{111}, \ldots X_{ijk}, \ldots, X_{12.4.3}, X_{Rest} \mid Z = n)$$

$$\sim Mult(n, \underbrace{p_1, \ldots}_{12\times}, p_2, \ldots\ldots, p_{12}, 1 - \sum_{i=1}^{12} 12p_i).$$

Here the probabilities p_i are given by the quotient of initial sample volume on the ith plate to the initial volume $p_i = \frac{v}{2^i V}$.
The following theorem establishes a link between our model and the Poisson model of Weiss et al.. Note that this is similar to a well-known result on the Poisson process, see e.g. Karlin and Taylor (1975, equation 2.3).

Theorem 1:

Let $Z, X_i, i = 1, \ldots, k$ random variables with $\sum_{i=1}^{k} X_i = Z$, $Z \sim P_{\lambda V}$ and $(X_1, \ldots, X_k \mid Z = n) \sim Mult(n, p_1, \ldots, p_k)$ with $\sum_{i=1}^{k} p_i = 1$.
Then the X_i are stochastically independent and Poisson-distributed random variables with $X_i \sim P_{p_i \lambda V}$.

Proof: Applying the theorem of total probabilities yields

$$
\begin{aligned}
P(\bigcap_{i=1}^{k}\{X_i = x_i\}) &= \sum_{n=0}^{\infty} P(\bigcap_{i=1}^{k}\{X_i = x_i\}, Z = n) \\
&= \sum_{n=0}^{\infty} P(\bigcap_{i=1}^{k}\{X_i = x_i\} \mid Z = n) \cdot P(Z = n)
\end{aligned}
$$

Since $P(\bigcap_{i=1}^{k}\{X_i = x_i\} \mid Z = n) = 0$ for all n with $n \neq \sum_{i=1}^{k} x_i$ we get

$$
P(\bigcap_{i=1}^{k}\{X_i = x_i\}) = P(\bigcap_{i=1}^{k}\{X_i = x_i\} \mid Z = \sum_{i=1}^{k} x_i) \cdot P(Z = \sum_{i=1}^{k} x_i).
$$

The common distribution of the $X_1, \ldots, X_k$ given $Z = n$ is multinomial with probabilities $p_1, \ldots, p_k$, whereas Z itself is Poisson-distributed. Therefore we obtain

$$
P(\bigcap_{i=1}^{k}\{X_i = x_i\}) = n! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!} \; \frac{(\lambda V)^n}{n!} e^{-\lambda V}.
$$

Thus, $n = \sum\limits_{i=1}^{k} x_i$ and $\sum\limits_{i=1}^{k} p_i = 1$ yield

$$
\begin{aligned}
P(\bigcap_{i=1}^{k}\{X_i = x_i\}) &= e^{-\sum_{i=1}^{k} p_i\lambda V} \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!} \lambda V^{\sum_{i=1}^{k} x_i} \\
&= \prod_{i=1}^{k} \frac{(p_i\lambda V)^{x_i}}{x_i!} e^{-p_i\lambda V} \\
&= \prod_{i=1}^{k} P(X_i = x_i).
\end{aligned}
$$

The last set of equations shows that the common distribution of the $X_i, i = 1, \ldots, k$ is a product of k independently Poisson-distributed variables with parameter $p_i\lambda V$. $\square$

In the GLP-experiment the volume of the samples v_N is equal to one millilitre and therefore we get $p_i = \frac{1}{2^i V}$ and $p_i\lambda V = \frac{1}{2^i}\lambda$. If in the multinomial model the additional assumption is made that the number of c.f.u. in the analysed material is a realisation of a Poisson-distributed random variable, the counts on the plates follow the same distribution as in the Poisson model of Weiss et al..

4 Simulations of the GLP-test under the multinomial model

Weiss et al. (1991) have discussed the statistical aspects of the GLP-test and suggested further evaluation of its performance by simulation studies. These studies should reflect errors occurring in daily laboratory work and model different levels of working quality. The generated data should serve

to characterise the GLP-test and especially to investigate the power of the homogeneity tests. We perform such studies for the multinomial model.
Dahms (1992, 1996) performed several simulation studies on the GLP-test. She generated colony counts according to a simulation design which tries to reflect the different working steps in performing the GLP-experiment. Dependencies of counts were modelled by using a dynamic structure of expected counts being specified during the simulation process (Dahms, 1996). Additionally, she modelled four main sources of working errors. These errors are also used in our simulation study, but we generate counts according to the multinomial model.
The "counting-error" reflects any mistakes occurring while counting the colonies on the plates. If during the process of pipetting incorrect quantities of material are transferred, this may be due to any of the following reasons:

- "Calibration errors", that are caused by inexact calibration scales on the pipettes.
- "Reading errors", describe an incorrect filling of the pipette by the analyst.
- "Draining errors", that occur if the pipette is not emptied completely and a rest of the material remains in the pipette.

These three errors are summarised as "pipetting-errors". We assume normal distributed calibration and reading errors. If draining errors occur, actual sampling volumes can only be too small and therefore this error is described by the absolute value of a normal distributed random variable:

$$\begin{aligned} \text{Calibration error} \quad & a && \sim N(0,\sigma_a^2), \\ \text{Reading error} \quad & b && \sim N(0,\sigma_b^2) \text{ and} \\ \text{Draining error} \quad & c && \text{with probability density} \end{aligned}$$

$$f_c(t) = \begin{cases} 0 & \text{for} \quad t \leq 0 \\ 2\varphi(\frac{x}{\sigma_c})\frac{1}{\sigma_c} & \text{for} \quad f > 0 \end{cases}$$

with $E(c) = \sigma_c$ and φ density of the standard normal distribution.

Pipetting errors produce deviations of the actual sample volume from the one prescribed by the GLP-test. Since errors are assumed to be independent their effects are additive. Simulated sample volumes are therefore the sum of the prescribed GLP-volumes and volumes produced by the errors:

$$\text{Simulated volume: } v = \text{ GLP-volume } + a + b - c.$$

To model the counting error colony counts generated in our simulation study are normally distributed random variables with expectation X_{ijk}^s:

$$\text{Simulated counts} \qquad X_{ijk}^z \sim N(X_{ijk}^s, \sigma_{count}^2).$$

Here, X_{ijk}^s denote the simulated number of c.f.u. on the ith evaluable dilution step in the jth series of the kth parallel plate proportional to the simulated volume described above. They are generated according to the conditional binomial distributions described in the section before.

In our simulation we study different combinations of these errors with various intensities, i.e. with different values for $\sigma_a^2, \sigma_b^2, \sigma_c^2$, and σ_{count}^2, reflecting different levels of working qualities. Values for the variances are based on

Table 1: Summary of simulated working errors (in size of standard deviation)

Working Errors	Run 0	Run 1	Run 2	Run 3
Pipetting of sample				
Calibration error w.r.t. target value	-	0.5 %	1 %	1 %
Reading error w.r.t. calibration mark	-	0.5 %	1 %	1 %
Draining error w.r.t. actual value	-	0.1 %	0.2 %	0.2 %
Pipetting of diluent fluid				
Calibration error w.r.t. target value	-	0.5 %	1 %	2 %
Reading error w.r.t. calibration mark	-	0.5 %	1 %	2 %
Draining error w.r.t. actual value	-	0.1 %	0.2 %	0.2 %
Counting error w.r.t. number of colonies	-	5 %	10 %	5 %

practical experience reported by Müller (1989). Note that the size of these variances is critical for the ensuing simulations. Any statement on the probability of the tests to falsely rejecting "good" work or falsely accepting "poor" work, heavily depends on these variances. First, a simulation run without any error terms is made (Run 0). These data provide material to assess the evaluation strategy of the GLP-test in case of ideal laboratory work. The second constellation of errors, Run 1, is supposed to reflect "good" laboratory work. The size of errors was chosen to describe small inaccuracies and mistakes which are unavoidable even for experienced analysts. Therefore, the GLP-test would be expected to accept these simulated data sets. In Run 2 the variance of the error terms was doubled to simulate more imprecise laboratory work producing data sets which should be rejected by the GLP-test. The last simulation run, Run 3, gives an impression of the influence of pipetting errors compared to counting errors. The exact values for the standard deviations used in the four simulation runs are presented in Table

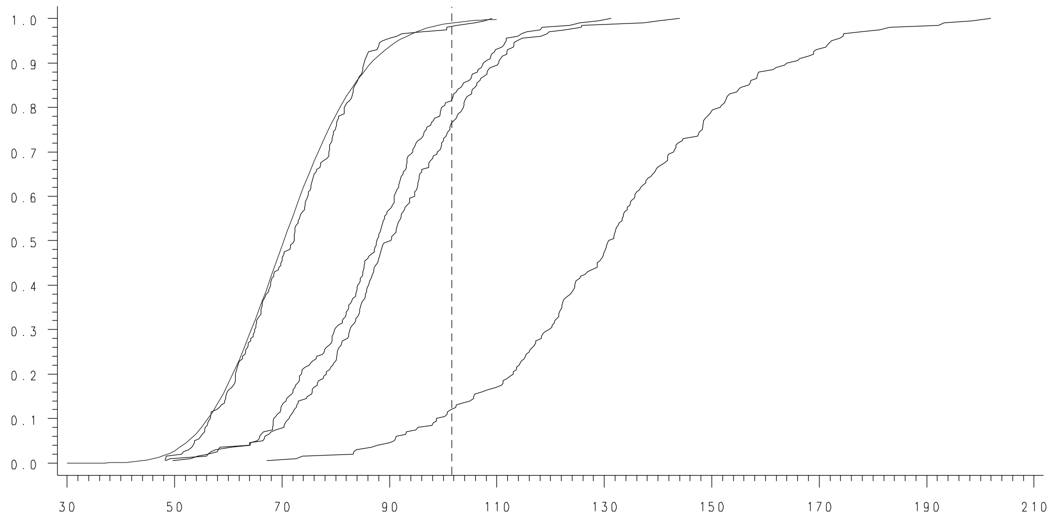

Figure 2: Simulated distribution functions of G_A^2- statistics compared to χ^2-distribution for 200 data sets (starting left: χ^2, Run 0, Run 1, Run 3, Run 2).

1.

For the simulations a germ concentration of $\lambda = 500$ per millilitre and a starting volume equal to $V = 30$ millilitre was assumed. Hence, the starting number of c.f.u. was $n = 15000$. Since in practice the sample material contains a fixed number of c.f.u., we decided to use a fixed number for n in the simulation study also. So we do not employ the assumption that this number n can be considered as a realisation of a Poisson-distributed variable. In each run 200 data sets were generated. Except for two data sets in Run 2 with five evaluable dilution levels, all data sets contained six evaluable levels. The simulated distributions of the G_S^2-statistic were comparable for all four runs. As described above, this test checks if colony counts of the parallel plates show less than the expected variation. This is not the case for any of the four runs and so this result was expected.

The simulated distributions of the G_A^2-statistic are displayed in Figure 2. For Run 1 and 3 containing the same counting error (5%) simulated distributions match closely. For Run 2 with the counting error twice as big (10%) the dis-

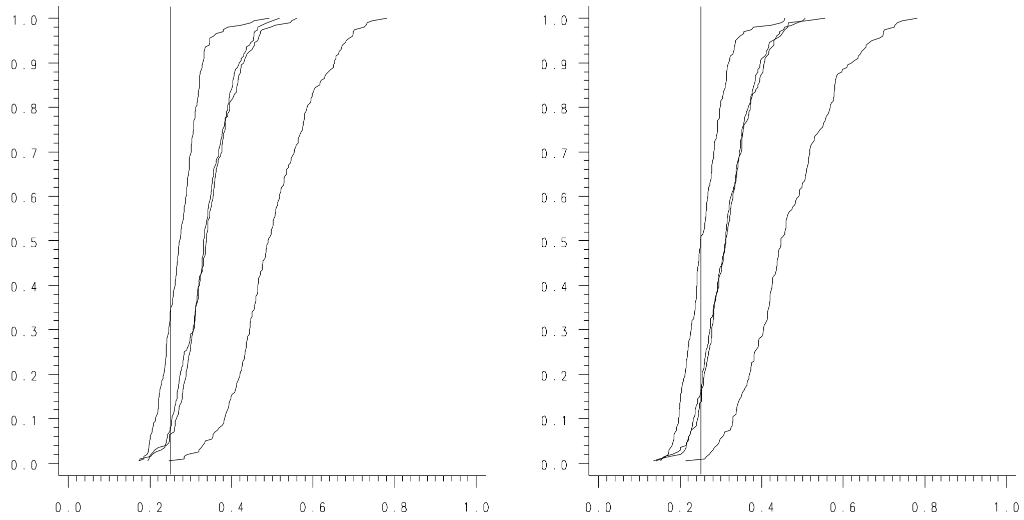

Figure 3: Simulated total variance (left) and variance between plates (right) for 200 data sets (starting left: Run 0, Run 1, Run 3, Run 2 and expected variance of 0.25).

tribution function is shifted to the right indicating that the G_A^2-statistic tends to much higher values. It can be seen very clearly that the influence of the counting error is much greater than the effect of the pipetting errors. In Run 1, describing "good" laboratory work, the test of overall homogeneity rejected nearly 19 % of the data sets as not meeting the required standard (see Table 2). Therefore, the G_A^2-test judges too strictly for the size of errors simulated here. The nominal level of significance $\alpha = 1\%$ was clearly exceeded. On the other hand, for Run 2, which modelled "poor" laboratory work, 11% of the data sets have not been rejected. The second part of the statistical analysis, based on the analysis of variance model described above,

Table 2: Assessments of the simulated error constellations

	Run 0	Run 1	Run 2	Run 3
G_S^2 : H_0 is rejected: repeat the experiment	0	0.5%	0	0
G_A^2 : H_0 is not rejected	98%	81.4%	11%	76.5%
G_A^2 : H_0 is rejected: 2. Evaluation step	2%	18.6%	89%	23.5%
Variance > 1: conclude "poor" performance	0	0	0	0

Table 3: Averages of the variance components

Source of variation	Run 0	Run 1	Run 2	Run 3
Total variance	0.276	0.336	0.503	0.344
Between dilution series	1.6%	1.9%	2.6%	3.4%
Dilution steps	4.8%	4.3%	4.6%	4.9%
Between plates	93.7%	93.8%	92.8%	91.8%

finally decides whether laboratory work meets the required standard. Here, neither in Run 1 nor in Run 2 any of the data sets was rejected. The total variance of the transformed data was clearly below the critical value 1 for all three simulated levels of working quality (see Table 3). The GLP-test accepts the simulated "poor" laboratory work for all generated 200 data sets as being acceptable laboratory work. The results of this simulations reveal a great discrepancy between the quality assessments made by the first and the second part of the statistical analysis.

The aim of the first part, containing the tests of homogeneity, is to avoid the time consuming analysis of variance components and to simplify the assessment of data quality (Dahms, 1992). This is not achieved for the working errors simulated here.

Since no data set of Run 2 is finally rejected, the critical value of 1 may be reduced. Table 4 shows the percentage of rejected data sets in the second part of the analysis for different critical values.

When using a critical value of 0.5 instead of 1 only 1% of the data sets in Run 1, describing "good" laboratory work, would be rejected while 47.5% of the data sets in Run 2 would be identified as "poor" laboratory work. Therefore, this value seems to be more suitable. Since the critical value of 1 is based only on practical experiences and does not have theoretical basis, a modification

Table 4: Critical values for the total variance

	Run 0	Run 1	Run 2
$G_A^2 : H_0$ rejected: 2. Evaluation step	2%	18.6%	89%
Variance > 1: repeat the experiment	0	0	0
Variance > 0.6: repeat the experiment	0	0	17%
Variance > 0.5: repeat the experiment	0	1%	47.5%
Variance > 0.4: repeat the experiment	2%	16%	85%

of this value might be a promising attempt to improve the performance of the test. When nearly the same decision should be made in part one and two of the analysis, in addition to reducing the critical value to 0.5 a modification in part one might also be considered (see also Figure 4). As described in Section 2 the test of overall homogeneity rejects $\mathcal{H}_0$ if G_A^2 is greater than $\chi^2_{IJK-1,0.99} = 101.62$, assuming 6 evaluable dilution levels. The value 101.62 was chosen to distinguish between ideal and non-ideal laboratory work but actually, we are looking for a critical value discriminating "good" from "poor" work. For simulation results obtained by us, a critical value of 126 might be useful (see Figure 4). With this value for Run 1, "good" laboratory work, only 1% of the data sets is rejected and in Run 2, 58% are identified as "poor" laboratory work (Figure 4). Nevertheless these critical values are very rough values based on relatively few simulations. If recommendation on this decision rule is desired, more exhaustive simulations have to be made.

5 Discussion

Statistical analysis of the data collected in the GLP-test as suggested by Weiss et al. is based on the assumption that observed colony counts are independent Poisson-distributed random variables. Based on simulation studies

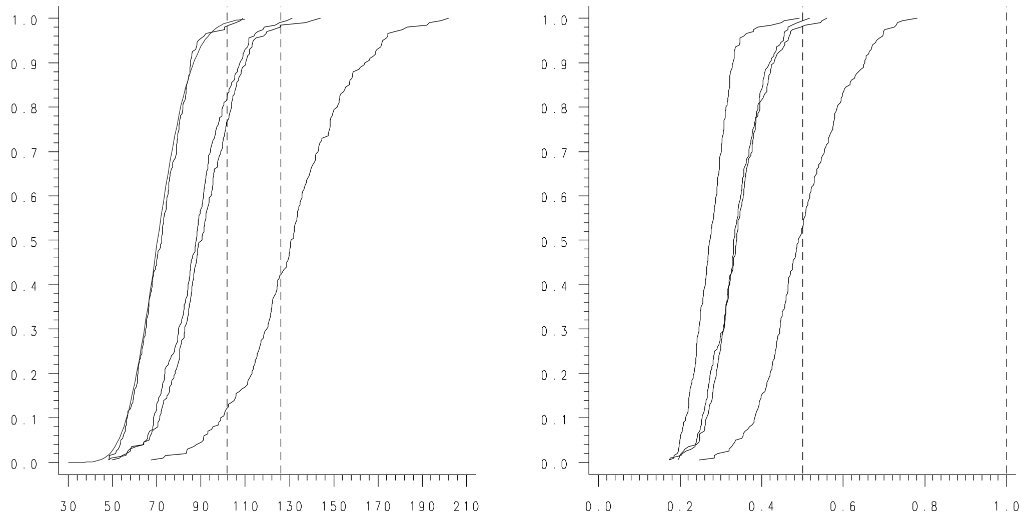

Figure 4: Simulated distribution functions of G_A^2 compared to χ^2-distribution with critical values 101.62 and 126 (left) and simulated total variance with critical values 0.5 and 1 (right)(starting left: Run 0, Run 1, Run 3, Run 2).

this assumption was criticised by Dahms (1996) who claimed that it simplifies reality and does not take into account certain dependencies inherent in the structure of the data.

In this article a new model for describing the distribution of the colony counts in the GLP- test is proposed. The model is based on the multinomial distribution. By introducing and combining conditional binomial distributions dependencies of samples were modelled and the limited number of c.f.u. in the analysed material was taken into account. Additionally, the multinomial model reflects the ratio of sample volumes to the volumes of the material where they are obtained from. We show that if in the multinomial model the additional assumption is made that the number of c.f.u. in the analysed material is a realisation of a Poisson-distributed random variable, then counts on the plates follow the same distribution as in the Poisson model of Weiss, Niemelä and Arndt (1991), i.e. they are independent Poisson-distributed variables. Based on the results of her simulation studies Dahms (1996) claims that the Poisson model does not take the dependency of c.f.u.in the GLP- test into account appropriately. Therefore, she strongly advises against the further use of the evaluation scheme in its current form. The multinomial model introduced in this paper closely reflects the dependencies inherent in the structure of the c.f.u.. The fact that this model is closely related to the Poisson model and our simulation results encourage further use of the Poisson model and the statistical evaluation scheme based on it.

The objectives of the GLP-test is to assess the quality of analyst performance. We investigated the effects of common working errors explored in the simulation studies of Dahms in the new context of the multinomial model. The results of our simulations reveal a great discrepancy between the quality assessments made by the first and the second part of the statistical analysis in the GLP- test. The objective of the first step seems not to be met, namely to simplify the analysis by avoiding the more complex second step.

In our simulations all data sets have been accepted as meeting the required standard of analysis, even those reflecting very "poor" working quality. Therefore, the critical value of 1 used by the GLP-test to decide over acceptance or rejection of a data set seems to be too high. For working errors simulated in this article a value of 0.5 appears to be more suitable. Since the critical value of 1 is based only on practical experiences and does not have sound theoretical basis, a modification of this value should be considered. The critical values derived by us provide a useful hint on how performance of the GLP-test may be improved. But simulation studies with more than 200 data sets per run should be performed to validate our suggestions.

One of the key problems of the simulation studies performed by Dahms and in this article is the assumption that variances of error terms used to model the different levels of working quality appropriately reflect the respective level of quality. If, for example, even experienced laboratory assistants produce errors higher than the ones modelled in Run 1, our judgement of the GLP-test would have been more positive. This issue requires further investigation.

References

BAUMGARD, J. (1986). *Mikrobiologische Untersuchung von Lebensmitteln*, Behr Verlag, Hamburg.

BUNDESGESUNDHEITSAMT (1991). *Bestimmung der Keimzahl; Verfahren zur Qualitätssicherung im Untersuchungslaboratorium*, Amtliche Sammlung von Untersuchungsverfahren nach §35 LMBG, Beuth Verlag, Berlin /Köln.

DAHMS, S. (1992). *Simulation als Mittel der Modellkritik*, Lit-Verlag.

DAHMS, S. (1996). *Simulation zur Bewertung mikrobiologischer Laborarbeit*, Biometrie und Informatik in der Medizin und Biologie **27**(1), p 32-47.

DIN ISO 5725 (1994). *Accuracy (trueness and precision) of measurement methods and results - Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method,* Beuth Verlag; Berlin/ Köln.

FARMILOE, F. J./ CORNFORD, S. J./ COPPOCK, J. P. M./ INGRAM, M. (1954). *The Survival of Bacillus Subtilis Spores in the Baking of Bread*, J. Sci. Food. Agric. **5**, p 292-304.

INTERNATIONAL DAIRY FEDERATION (1994). *Quality Control in the Microbiological Laboratory: Analyst Performance Assessment for Colony Counts*, IDF-Standard, p 169.

ISO/CD 14461 (1996). *Milk and Milkproducts - Quality control in the Microbiological Laboratory - Analyst Performance Assessment for Colony Count.*

Karlin, S./ Taylor, H. (1975). *A First Course in Stochastic Processes*, Academic Press, Inc..

Kotz S./ Johnson N. L. (1981). *Encyclopedia of Statistical Science*, Wiley & Sons, New York.

Müller, A. (1989). *Die Bedeutung des Stichproben- und Methodenfehlers für die kulturelle Keimzahlbestimmung*, Dissertation, Fachbereich Veterinärmedizin, Freie Universität Berlin.

Schlittgen, R. (1996). *Statistische Interferenz*, R. Oldenbourg Verlag, München.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*, Wiley & Sons, New York.

Weiss, H./ Niemelä, S./ Arndt, G. (1991). *Sicherung der Präzision standardisierter mikrobiologischer Untersuchungsverfahren*, Biometrie und Informatik in der Medizin und Biologie **22**(3), p 116-135.